Supplementary File S1

"Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research"

À. Bravo, J. Piñero, N. Queralt, M. Rautschka, L.I. Furlong.

Evaluation of the RE system on the AIMED corpus

AIMed corpus

The AIMed corpus is widely used for PPI extraction (ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/). The AImed corpus consists of 225 MEDLINE abstracts, of which 200 abstracts describe interactions between human proteins and 25 do not refer to any interaction. There are 5625 annotated sentences, 1008 containing a true PPI (TRUE) and 4617 not containing a true PPI (FALSE).

Evaluation of Kernel based Relation Extraction

The performance of each model for association classification was evaluated by sentence-level 10-fold cross validation in each corpus. The classifiers' performances were assessed using P, R and F-score over the class TRUE. TRUE sentences contain real relationship between the entities analysed, in contrast with FALSE sentences where the two entities co-occur, but there is no semantic relationship between them.

Results

The first experiments were conducted on the AIMed corpus in order to evaluate the performance of the K_{DEP} kernel with different features and compare it with previous results (Table 1). Compared to the performance obtained with the original system (K_{SL} with sparse bigrams, P 52.9%, R 65.3%, F 58.1 %), K_{DEP} alone does not show an improvement in performance (26: P 41.3%, R 61.2%, F 49%). However, the combination of K_{SL} and K_{DEP} kernels results in an improvement of the results both at the level of precision and recall. For example, using role, lemma and stem as features leads to high levels of F-score (96: P 55.5%, R 67.2%, F 60.3%; 98: P 55.3%, R 67.3%, F 60.3%). Although we show experiments where the features on the token for the v-walk and e-walk are tested one at a time, we also conducted experiments testing different combinations of these features, but the results were not better than the ones obtained with single features over the token (results not shown). Thus, the use of shallow linguistic information but also syntactic information in the form of dependency walk features lead to more accurate models for PPI relation extraction. Our results are comparable to the results obtained with state-of-the-art approaches tested on the AIMED corpus [1].

Table 1: Selected results obtained by 10-fold cross-validation on the AIMed corpus.
The best results achieved are highlighted in bold. The first column indicates the number of the experiment, the second column shows if K_{SL} is used with (SB)/without (O) sparse bigrams, or if it is not used (-). The next two columns focus on K_{DEP} features, that can be represented as v-walk and/or e-walk: token (T), stem (S), lemma (L), POS-tag (P), role (R) or none (-). The last columns show the result obtained in each experiment indicating precision (P), recall (R) and f-measure (F1) in percentage (%).

Num.	K_{SL}	K_{DEP}		AIMed		
		v-walk	e-walk	P	R	F1
1	O	-	-	51.4	65.9	57.4
2	SB	-	-	52.9	65.3	58.1
6	-	R	-	35.7	63.2	45.3
26	-	R	L	41.3	61.2	49.0
36	-	R	T	41.5	60.5	48.8
41	O	R	-	52.6	69.2	59.3
96	SB	R	L	**55.5**	**67.2**	**60.3**
98	SB	S	R	55.3	67.3	60.3
103	SB	S	T	57.4	62.3	59.4

References

1. Chowdhury MFM, Lavelli A: **Combining tree structures, flat features and patterns for biomedical relation extraction**. In *EACL '12 Proc 13th Conf Eur Chapter Assoc Comput Linguist*. Association for Computational Linguistics; 2012:420–429.